A Study of Defensive Methods to Protect Visual Recommendation Against Adversarial Manipulation of Images

Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, Felice Antonio Merra*
name.surname@poliba.it
Polytechnic University of Bari
Bari, Italy

ABSTRACT

Visual-based recommender systems (VRSs) enhance recommendation performance by integrating users' feedback with the visual features of items' images. Recently, human-imperceptible image perturbations, defined *adversarial samples*, have been shown capable of altering the VRSs performance, for example, by pushing (promoting) or nuking (demoting) specific categories of products. One of the most effective adversarial defense methods is *adversarial training* (AT), which enhances the robustness of the model by incorporating adversarial samples into the training process and minimizing an adversarial risk. The AT effectiveness has been verified on defending DNNs in supervised learning tasks such as image classification. However, the extent to which AT can protect deep VRSs, against adversarial perturbation of images remains mostly under-investigated.

This work focuses on the defensive side of VRSs and provides general insights that could be further exploited to broaden the frontier in the field. First, we introduce a suite of adversarial attacks against DNNs on top of VRSs, and defense strategies to counteract them. Next, we present an evaluation framework, named Visual Adversarial Recommender (VAR), to empirically investigate the performance of defended or undefended DNNs in various visually-aware item recommendation tasks. The results of large-scale experiments indicate alarming risks in protecting a VRS through the DNN robustification. Source code and data are available at https://anonymous.4open.science/r/503dde32-af4c-4e29-8e55-2a908f57e64b/.

KEYWORDS

Adversarial Machine Learning; Recommender System; Multimedia Recommendation

ACM Reference Format:
Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, Felice Antonio Merra. 2021. A Study of Defensive Methods to Protect Visual Recommendation Against Adversarial Manipulation of Images. In *Proceedings of ACM Conference (SIGIR'21).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*The authors are in alphabetical order. Corresponding author: Felice Antonio Merra (felice.merra@poliba.it), Daniele Malitesta ((daniele.malitesta@poliba.it)).

SIGIR'21, July 2021, Online Event (pre-print)

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recommender systems (RSs) have terrifically taken over online shopping by providing users with personalized recommendations to disentangle the chaotic flood of products on e-commerce platforms. RSs model the complex preference that consumers exhibit toward items by leveraging a sufficient amount of past behavioral data. Accordingly, in scenarios such as fashion, food, or point-of-interest recommendation, images associated with products can impact the outcomes of purchasing/consumption decisions, as images attract attention, stimulate emotion, and shape users' first impression about products and brands. To extend the expressive power of RSs, visual-based recommender systems (VRSs) have recently merged that attempt to incorporate products' visual appearance of items into the design space of RS models [14]. Given the representational power of deep neural networks (DNNs) in capturing characteristics and semantics of the images, state-of-the-art VRSs often incorporate visual features extracted via a DNN — pre-trained, e.g., VBPR [23] and ACF [10], or learned end-to-end, e.g., DVBPR [26] — and integrate it with a recommendation model such as matrix factorization (MF) to better judge the users' interests.

For instance, He and McAuley [23] proposed VBPR, the pioneering visual-aware MF method based on BPR [42] that integrates visual features extracted from a pre-trained DNN, yielding superior performance over the baseline version of the same recommender, i.e., BPR-MF. Chen et al. [10] proposed ACF by modeling component- and item-level image representations via two attention networks where the first network learns the users' interest toward different regions of the product image and the second network learns to score an unseen product comparing it with the interacted ones.

Despite their great success, DNNs have been found vulnerable to adversarial examples [13], which means very small changes to the input image can fool a pre-trained DNN to misclassify the adversarial image with high confidence. There is now a sizable body of work proposing different attack and defense strategies in adversarial setting, namely FGSM [18], PGD [33], and Carlini & Wagner [8] (for the attacks), and Adversarial Training [18], Free Adversarial Training [45] (on the defensive side). These works constitute essential first steps in exploring the realm of adversarial machine learning (AML), i.e., machine learning (ML) in an adversary's presence.

Research in the AML field has evolved significantly over the last eight years and beyond, from the pioneering work on the security of ML algorithms by Szegedy et al. [46] to more recent works in the context of object detection [40], malware detection [55], speech recognition [25], graph learning [16], and adversarial attacks against item recommendation task [24]. As for the latter, recently He et al. [24] demonstrated the weakness of BPR-MF recommenders with respect to adversarial perturbations on model embeddings and proposed an adversarial training (AT) procedure to make the resultant model more robust. Similarly, Tang et al. [47] verified the efficacy of AT in protecting VBPR against perturbations applied directly on the image features extracted via ResNet50 [21]. Moreover, Di Noia et al. [15] demonstrated that targeted adversarial attacks on input images —and not on their features as studied in [47]— can maliciously alter the recommendation performance.

Notwithstanding these efforts, research in the AML-RS field has been predominately focused on attacks on and defense of collaborative filtering (CF) models, such as BPR-MF [24], collaborative auto-encoder [53], tensor-factorization [9], and self-attention sequential [34] models. However, generating adversarial images that are similar to source images (via pixel-level perturbations) and are capable of comprising the quality of VRSs, makes the attacks much stronger from a practical perspective since the depicted scenario is more realistic. Imagine the following motivational example: a competitor is willing to increase the recommendability of a category of products on an e-commerce platform, e.g., *sandals*, for economical gain. She can achieve this goal by simply uploading adversarially perturbed product images of sandals that are misclassified by the DNN used in the VRS, named image feature extractor (IFE), as a much more popular class, e.g., *running shoes*, allowing sandal to be pushed into recommendation list of more users.

The work at hand focuses on discovering the unknown vulnerability of VRSs against the poisoning of training data with adversarially-perturbed product images constructed to be misclassified by the IFE. In this respect, we propose an empirical framework, named *Visual Adversarial Recommendation* (VAR), to study whether and to what extent adversarial training strategies can strengthen IFE's classification performance, thus mitigating the adverse effects of such attacks on the recommendation task. Then, we study whether the class of VRSs that internally trains the IFE, e.g., DVBPR [26], could be still affected by adversarial samples crafted on a pretrained DNN, e.g., ResNet50, and *transferred* to this end-to-end class of VRSs.

The main contributions of this work are summarized as follows:

(1) an extensive study of adversarial training (defensive) methods to robustify the visually-aware recommendation performance through the analysis of 156 combinations of three type of IFEs, three attacks, and five VRSs, and three recommendation datasets;

(2) the proposal of a novel rank-based evaluation metric, named *category normalized Discounted Cumulative Gain*;

(3) analysis of the variation of global and beyond-accuracy recommendation performance with (and without) defenses to understand to what extent the adversaries in our VAR setting are altering the overall performance of the recommender.

The rest of the paper is organized as follows. First, we present the framework in Section 2. In Sections 3 and 4 we introduce the experimental setups and present and discuss the empirical results. Finally, we report the related work in Section 5 and we present conclusions and possible future directions in Section 6.

2 THE PROPOSED FRAMEWORK

In this section, we describe the proposed Visual Adversarial Recommendation (VAR) experimental framework. First, we define some preliminary concepts. Then, we provide an overview of all VAR components. Finally, we present the evaluation measures to quantify the effectiveness of the adversarial defenses under attacks.

2.1 Preliminaries

Recommendation Task. We define the set of users, items and 0/1-valued preference feedback as $\mathcal{U}$, $\mathcal{I}$, and $\mathcal{S}$, where $|\mathcal{U}|$, $|\mathcal{I}|$, and $|\mathcal{S}|$ are the set sizes respectively. The preference of a user $u \in \mathcal{U}$ on item $i \in \mathcal{I}$ is encoded with $s_{ui} \in \mathcal{S}$, in which we assume that the user likes the item (i.e., $s_{ui} = 1$), if she has interacted (i.e., reviewed, purchased, clicked) with the item. Furthermore, we define the recommendation task as the action to suggest items that maximize, for each user, a utility function. We indicate with $\hat{s}_{ui}$ the predicted score learned from the recommender system (RS) upon historical preferences, represented as a user-item preference-feedback matrix (UPM), which is usually *high-dimensional* and *sparse.* Matrix Factorization (MF) [28] trains a model to approximate the UPM as the dot product of two much smaller embeddings, i.e., a *user latent* vector $p_u \in \mathbb{P}^{|\mathcal{U}|\times h}$, and an *item latent* vector $q_i \in \mathbb{Q}^{|\mathcal{I}|\times h}$, where $h << |\mathcal{U}|, |\mathcal{I}|$.

Deep Neural Network. Given a set of data samples (x_i, y_i), where x_i is the i-th image and y_i is the one-hot encoded representation of x_i's image category, we define F as a DNN classifier function trained on all (x_i, y_i). Then, we set $F(x_i) = \hat{y}_i$ as the predicted probability vector of x_i belonging to each of all the admissible output classes, and we calculate its predicted class as the index of $\hat{y}_i$ with maximum probability value, and represent it as $F_c(x_i)$. Moreover, assuming an L-layers DNN classifier, we indicate with $F^{(l)}(x_i)$, $0 \leq l \leq L-1$, the output of the l-th layer of F given the input x_i.

Adversarial Attack and Defense. We define an *adversarial attack* as the problem of finding the best value for a perturbation δ_i such that (i) the attacked image $x_i^* = x_i + \delta_i$ must be visually similar to x_i according to a certain *distance* metric, e.g., L_p norms, (ii) the predicted class for x_i^* must be different from the original one, i.e., $F_c(x_i + \delta_i) \neq F_c(x_i)$, and (iii) x_i^* must stay within its original value range, i.e., $[0, 1]$ for 8-bit RGB images re-scaled by a factor 255. When $F_c(x_i^*)$ is required to be generically different from $F_c(x_i)$, we say the attack is *untargeted.* On the contrary, when $F_c(x_i^*)$ is specifically required to be equal to a target class t, we say the attack is *targeted.* Finally, we define a *defense* as the problem of finding ways to limit the impact of adversarial attacks against a DNN. For instance, a standard solution consists of training a more robust version of the original model function —we will refer to it as $\widetilde{F}$— which attempts to correctly classify attacked images.

2.2 Empirical Framework

Here, we describe the VAR components shown in Figure 1: *adversary*, *image feature extractor*, and *visual-based recommender system.*

Adversary. To align with the AML literature, we follow the attack —and defense— adversary threat model outlined in [5]. Given all the top-K recommendation lists generated by the VRS, the *adversaries' goal* is to push the items at the bottom of the lists to higher

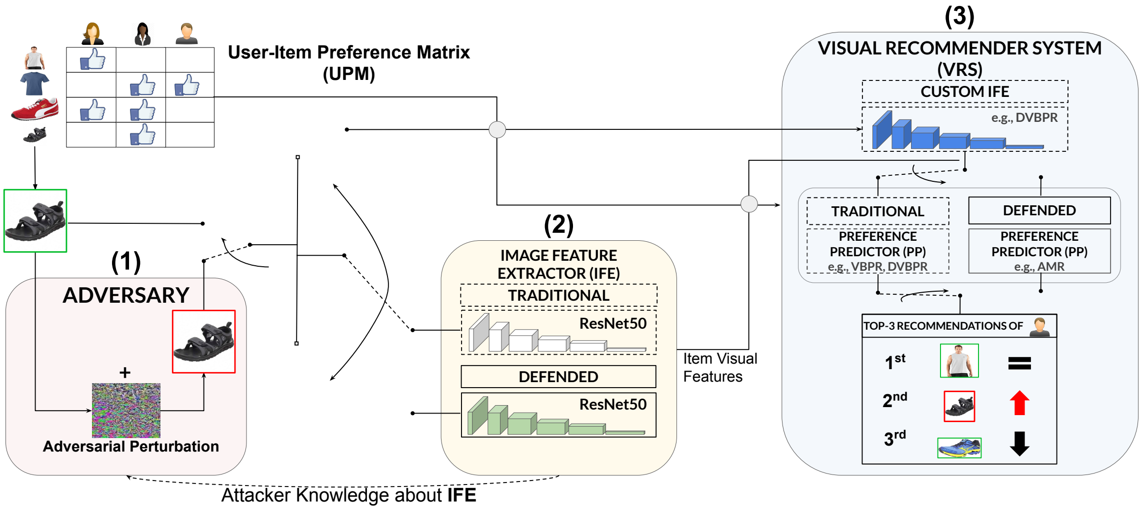


Figure 1: Overview of our VAR framework. (1) an *Adversary* might perturb product images. (2) an *Image Feature Extractor* (IFE) extracts the item visual features. The IFE is implemented either with an external, pre-trained DNN or with a custom DNN within the *Visual Recommender Systems* (VRS). (3) the *Preference Predictor* (PP) from the VRS takes the user-item preference matrix (UPM) and the visual features to compute the top-K lists. Adversarial training strategies can protect both the external IFE and/or the PP.

positions. We assume that adversaries are aware of recommendation lists and choose the low-ranked category of item to be pushed (*source*). Then, they select the category of a more recommended item (*target*). Two additional assumptions arise here: *(1)* the adversaries have perfect knowledge of the image feature extractor (IFE) used in the VRS and perturb source images to be mis-classified as target ones, i.e., *white-box* attack setting, which is the *worst-case attack scenario*, or *(2)* they cannot access the IFE, since it is end-to-end trained along with the VRS, and craft the adversarial samples on another DNN to be transferred on the victim's recommender, i.e., *black-box* attack setting. In our motivating scenario, the adversaries can *poison* the dataset by uploading the adversarially corrupted item images on the platform that employs a VRS.

Image Feature Extractor (IFE). The input sample x_i represents the photo associated with the item $i \in \mathcal{I}$, which may appear in the top-K recommendation list shown to a user. Hence, the IFE is a DNN to extract high-level visual features from x_i. The model can be either *pretrained* on a classification task, i.e., He et al. [21], or a custom network trained *end-to-end* along with the VRS, i.e., Kang et al. [26]. The actual extraction takes place at one of the last layers of the network, i.e., $F^{(e)}(x_i)$, where e refers to the extraction layer. In general, we define this layer output $F^{(e)}(x_i) = \varphi_i$ as a three-dimensional vector that will be the input to the VRS. No defense is applied on the custom IFE (see Figure 1) used in DVBPR since defensive approaches only refer to *classification* models. Note that the IFE is a key component in VAR since it represents the connection between the adversary —responsible for the attack— and the VRS.

Visual-based Recommender System (VRS). In VAR, the VRS is the component aimed at addressing the recommendation task. The model takes two inputs: (i) the historical UPM, and (ii) the set of item visual features extracted from the pretrained IFE or custom IFE, i.e., DVBPR [26]. Thus, it produces recommendation lists sorted by the preference prediction score evaluated for each user-item pair. Indeed, the VRS preference predictor takes advantage of the pure collaborative filtering source of data, i.e., the UPM, and the high-level multimedia features to unveil user's preferences [23]. In the VAR motivating example, the VRS is the final victim of the adversary. For this reason, in this work we focus our analysis on the performance variation of the VRS, both in the attack and defense scenarios. The final objective is to analyze the robustness/vulnerability of different VRSs, that are being influenced by different settings of the adversary and the IFE.

2.3 Evaluation

We perform three levels of investigation, namely: (i) the effectiveness of adversarial attacks in misusing the classification performance of the DNN used as the IFE, (ii) the variation of the accuracy— and beyond-accuracy— recommendation performance, and (iii) the evaluation of consequences for attack and defense mechanisms on the recommendability of the category of items to be pushed.

In AML, several publications focused on quantifying adversarial attacks success in corrupting the classification performance of a target classifier, i.e., the attack Success Rate (*SR*) [8]. Similarly, there is a vast literature about the accuracy and beyond accuracy of RSs [44] recommendation metrics. On the other hand, we have observed a lack of literature evaluating adversarial attacks on RSs content data. As a matter of facts, Tang et al. [47] evaluate the effects of untargeted attacks on classical system accuracy metrics, i.e., *Hit Ratio* (*HR*) and *normalized Discounted Cumulative Gain* (*nDCG*), while Di Noia et al. [15] propose a modified version of *HR* to evaluate the fraction of adversarially perturbed items in the top-K recommendations. To fill this gap, we redefine the Category Hit Ratio (*CHR@K*) [15] and formalize the normalized Category Discounted Cumulative Gain (*nCDCG@K*).

Definition 1 (Category Hit Ratio). *Let C be the set of the classes extracted from the IFE, and $\mathcal{I}_c = \{i \in \mathcal{I}, c \in C | F_c(x_i) = c\}$*

be the set of items whose images are classified by the IFE in the c-class, e.g., the category of low recommended items. Then, we define categorical hit (chit) as:

$$chit(u,k) = \begin{cases} 1, & \text{if } k\text{-th item in the top-}K \in \mathcal{I}_c \\ 0, & \text{if } k\text{-th item in the top-}K \notin \mathcal{I}_c \end{cases} \tag{1}$$

where categorical hit ($chit(u,k)$) is a 0/1-valued function that is 1 when the item in the k-th position of the top-K recommendation list of the user u is in the set of attacked items not-interacted by u. Consequently, we define the CHR@K as follows:

$$CHR_u@K = \frac{1}{K}\sum_{k=1}^{K} chit(u,k) \tag{2}$$

Since *Category Hit Ratio* does not pay attention to the ranking of recommended items, we propose a novel rank-wise positional metric, named **Category normalized Discounted Cumulative Gain**, that assigns a *gain* to each considered ranking position. By considering a relevance threshold τ, we assume that each item $i \in \mathcal{I}_c$ has an ideal relevance value of:

$$idealrel(i) = 2^{(s_{max}-\tau+1)} - 1 \tag{3}$$

where s_{max} is the maximum possible score for items. By considering a recommendation list provided to user u, we define the relevance ($rel(\cdot)$) of a suggested item i as:

$$rel(k) = \begin{cases} 2^{(s_{ui}-\tau+1)} - 1, & \text{if } k\text{-th item} \in \mathcal{I}_c \\ 0, & \text{if } k\text{-th item} \notin \mathcal{I}_c \end{cases} \tag{4}$$

where k is the position of the item i in the recommendation list. In Information Retrieval, the *Discounted Cumulative Gain* (*DCG*) is a metric of ranking quality that measures the usefulness of a document based on its relevance and position in the result list. Analogously, we define Category Discounted Cumulative Gain (*CDCG*) as:

$$CDCG_u@K = \sum_{k=1}^{K} \frac{rel(k)}{\log_2(1+k)} \tag{5}$$

Since recommendation results may vary in length depending on the user, it is not possible to compare performance among different users, so the cumulative gain at each position should be normalized across users. In this respect, we define the *Ideal Category Discounted Cumulative Gain* (*ICDCG@K*) as follows:

$$ICDCG@K = \sum_{k=1}^{min(K,|\mathcal{I}_c|)} \frac{rel(k)}{\log_2(1+k)} \tag{6}$$

In practical terms, *ICDCG@N* indicates the score obtained by an ideal recommendation list that contains only relevant items.

Definition 2 (normalized Category Discounted Cumulative Gain). *Let C be the set of the classes extracted from the IFE, $\mathcal{I}_c = \{i \in \mathcal{I}, c \in C | F_c(x_i) = c\}$ be the set of items whose images are classified by the IFE in the c-class, i.e., the category of low recommended items. Let $rel(k)$ be a function computing the relevance of the k-th item of the top-K recommendation list, and ICDCG@K be the CDCG for an ideal recommendation list only composed of relevant items. We define the normalized Category Discounted Cumulative Gain (nCDCG), as:*

$$nCDCG_u@K = \frac{1}{ICDCG@K}\sum_{k=1}^{K} \frac{rel(k)}{\log_2(1+k)} \tag{7}$$

The *nCDCG@K* is ranged in a $[0,1]$ interval, where values close to 1 mean that the attacked items are recommended in higher positions, e.g., the attack is effective. In Information Retrieval, a logarithm with base 2 is commonly adopted to ensure that all the recommendation list positions are discounted.

3 EXPERIMENTAL SETUP

In this section, we first introduce the three real-world datasets, the adversarial attack strategies, the defense methods to make the IFE more robust, and the VRSs. Then, we present the complete set of evaluation measures and a detailed presentation of the experimental choices to make the results reproducible.

3.1 Datasets

Amazon Women & Amazon Men [22, 23, 35] are two datasets about men's and women's clothing from the Amazon category "Clothing, Shoes and Jewelry". Once having downloaded the images with a valid URL, we applied k-core filtering first on users and then on items to reduce the impact of cold users and items, as suggested by Rendle *et al.* [43]. While for `Amazon Men` we run 5-core filtering as suggested in [22, 23], for `Amazon Women` we adopted 10-core filtering to reduce its higher number of user/item interactions, and so reducing the VRS training time and the expensive hardware computation time in crafting adversarially perturbed product images [54]. This pre-processing step produced the following statistics: `Amazon Women` counts 54,473 interactions recorded between 16,668 users and 2,981 items, while `Amazon Men` count 89,020 interactions recorded 24,379 users and 7,371 items. **Tradesy** [23] dataset contains implicit feedback, i.e., purchase histories and desired products, from the homonym second-hand selling platform. We applied the same pre-processing pipeline described above. As for `Amazon Women`, we run *10*-core filtering. The final dataset counts 21,533 feedback recorded on 6,253 users and 1,670 products.

3.2 Adversarial Attacks and Defenses

In this section, we present all the adversarial attack and defense techniques adopted in the experimental phase.

3.2.1 Attacks. We explored three state-of-the-art adversarial attacks against DNNs image classifiers.

Fast Gradient Sign Method (FGSM) [18] is an L_∞-norm optimized attack that produces an adversarial version of a given image in just one evaluation step. A perturbation budget ϵ is set to modify the strength —and consequently, the visual perceptibility— of the attack, i.e., higher ϵ values mean stronger attacks but also more evident visual artifacts.

Projected Gradient Descent (PGD) [33] is an L_∞-norm optimized attack that takes a uniform random noise as the initial perturbation, and *iteratively* applies an FGSM attack with a continuously updated small perturbation α —clipped within the ϵ-ball— until either it effectively reaches the network misclassification, i.e.,

$F_c(x_i + \alpha_i) = t$, or it completes the number of possible iterations, i.e., 10 iterations in our evaluation setting.

Carlini and Wagner attacks (C & W) [8] are three attack strategies based on L_0, L_2 and L_∞ norms that re-formulate the traditional adversarial attack problem (see Section 2.1) by replacing the distance metric with a well-chosen *objective function*. C & W integrates the parameters κ, i.e., the *confidence* of the attacked image being classified as t, and a, i.e., the trade-off between optimizing the objective function and the classifier loss function.

3.2.2 Defenses. We explored two defense strategies.

Adversarial Training (AT) [18] consists of injecting adversarial samples into the training set to make the trained model robust to them. The major limitations of this idea are that it increases the computational time to complete the training phase, and it is deeply dependent on the type of attack strategy used to craft adversarial samples. For instance, Madry *et al.* [33] generates adversarial images with the PGD-method to make the trained model robust against both one-step and multi-step attack strategies.

Free Adversarial Training (FAT) [45] proposes a training procedure 3 – 30 times faster than the classical Adversarial Training [18, 33]. Unlike the previous one, this method updates both the model parameters and the adversarial perturbations doing a unique backward pass in which gradients are computed on the network loss. Moreover, to simulate a multi-step attack —which would make the trained network more robust— it keeps retraining on the same minibatch for m times in a row.

3.3 Visual-based Recommender Models

To evaluate VAR approach, we considered five VRSs. Table 1 presents an overview of the IFE components of the tested VRSs.

Factorization Machine (FM) [41] is a recommender model proposed by Rendle [41] to estimate the user-item preference score with a factorization technique. For a fair comparison with VBPR and AMR, we used BPR [42] loss function to optimize the personalized ranking. In this respect, we adopted LightFM [30] implementation integrating the UPM with the extracted continuous features. It is worth noticing that, differently from the recommenders we will present later, this model is not specifically designed for visually-aware recommendation tasks.

Visual Bayesian Personalized Ranking (VBPR) [23] improves the MF preference predictor by adding a *visual* contribution to the traditional *collaborative* one. Given a user u and a non-interacted item i, the predicted preference score is $\hat{s}_{ui} = p_u^T q_i + \theta_u^T \theta_i + \beta_{ui}$, where $\theta_u \in \Theta^{|\mathcal{U}| \times v}$ and $\theta_i \in \Theta^{|\mathcal{I}| \times v}$ are the *visual* latent vectors of user u and item i respectively ($v << |\mathcal{U}|, |\mathcal{I}|$). The visual latent vector of item i is obtained as $\theta_i = \mathbf{E}\varphi_i$, where φ_i is the visual feature of image item i extracted from a pretrained AlexNet [29] and $\mathbf{E}$ is a matrix to project the visual feature into the same space as of θ_u. Furthermore, β_{ui} includes the sum of the overall offset, and the user, item and global visual bias.

Attentive Collaborative Filtering (ACF) [10] tries to unveil the *implicitness* of multimedia user/item interactions by means of two *attention* networks. That is, one network learns to weight each user's interacted, i.e., positive items —because they are not equally *important* to the user— while another network learns to weight each *component* of the *feature map* extracted from the product image within the interacted items, e.g., regions of an image or frames of a video. Given a user u and a non-interacted item i, the predicted preference score is $\hat{s}_{ui} = \left(p_u + v_u\right)^T q_i$, where $v_u \in \mathbb{V}^{|\mathcal{U}| \times h}$ is an additional *user latent* vector weighted by the two *attention*-levels, i.e., item and component, described above.

Table 1: Tested VRSs. ***(FC: Fully-Connected, FM: Feature Maps)***

VRS		Image Feature Extractor			
		Extraction Layer		Training	
Model	*Reference*	*FC*	*FM*	*Pretrained*	*End-to-End*
FM	Rendle [41]	✓		✓	
VBPR	He and McAuley [23]	✓		✓	
AMR	Tang et al. [47]	✓		✓	
ACF	Chen et al. [10]		✓	✓	
DVBPR	Kang et al. [26]	✓			✓

Visually-Aware Deep BPR (DVBPR) [26] enhances the preference predictor proposed by He and McAuley [23] by replacing the pretrained visual feature extractor with a custom Convolutional Neural Network (CNN), which is trained *end-to-end* together with the preference predictor on the main recommendation task. Given a user u and a non-interacted item i, the predicted preference score is $\hat{s}_{ui} = \theta_u^T F^{(e)}(x_i)$, where θ_u is the user visual profile seen for VBPR and F is the custom CNN.

Adversarial Multimedia Recommendation (AMR) [47] is an extension of VBPR that integrates the adversarial training procedure proposed by He *et al.* [24] named *adversarial regularization* to build a model that is increasingly robust to FGSM-based perturbations against image features. Apart from the different training procedures, the score prediction function is the same as VBPR.

3.4 Evaluation Metrics

In addition to *CHR* and *CnDCG* shown in Section 2, we studied both the consequences of adversarial images on the IFE and variation of overall recommendation performance.

Adversarial attacks and defenses performance are evaluated through the attack Success Rate (*SR*) and the Feature Loss (*FL*), i.e., the mean squared error between the extracted image features before and after the attack.

Recommendation performance is evaluated with *Recall@K*, that is an accuracy metric that considers the fraction of recommended products in the top-K recommendation that hit test items, and the expected free discovery (*EFD@K*), a beyond-accuracy metric that provides a measure of the ability of an RS to recommend relevant long-tail items [49]. Since we are interested in measuring whether the application of targeted adversarial attacks might alter the overall performance of the RS, Table 5 reports the percentage variation of the performance between the attacked recommender and the base one. The reported metric is evaluated as follow

$$\Delta_{Rec} = \frac{\frac{1}{|Attacks|}\left(\sum_{a \in Attacks} Rec_a\right) - Rec_{Base}}{Rec_{Base}} \times 100 \qquad (8)$$

where *Attacks* indicates the set of tested attacks, e.g., FGSM, PGD, and C & W, and *Base* indicates that the metric value has been computed on the not-attacked recommender. The same formulation has been used to evaluate the Δ_{EFD}. Note that Δ negative values indicate a reduction of the performance.

3.5 Reproducibility

Adversarial attacks. Attacks were implemented with the Python library CleverHans [38]. For both FGSM and PGD, we adopted $\epsilon = 4$ re-scaled by 255. Then, for PGD's α parameter, we set the multi-step size as $\epsilon/6$ and the number of iterations to 10. As for the C&W attack, we ran a 5-step binary search to calculate a, starting from an initial value of 10^{-2} and set κ to 0. Furthermore, we set the maximum number of iteration to 1000 and adopted Adam optimizer with a learning rate of 5×10^{-3} as suggested in C&W [8]. Note that, to reproduce a real attack scenario, we saved the adversarial images in `tiff` format, i.e., a lossless compression, as lossy compression, e.g., JPEG, may affect the effectiveness of attacks [20].

Feature extraction. We used the PyTorch pretrained implementation of ResNet50 [21] to extract high-level image features. For FM, VBPR, and AMR, we set `AdaptiveAvgPool2d` as extraction layer, whose output is a 2048-dimensional vector. For ACF, we set the last `Bottleneck` output, i.e., its final `relu` activation, as extraction layer, whose output is a $7 \times 7 \times 2048$-dimensional vector. Finally, for DVBPR, we reproduced the exact same CNN architecture described in the original paper [26], whose extraction layer output is a 100-dimensional vector. Here, we adopted TensorFlow.

Defenses. In the non-defended scenario, we adopted ResNet50 pre-trained on `ImageNet` with traditional training. On the other hand, we adopted ResNet50 pre-trained on `ImageNet` with Adversarial Training and Free Adversarial Training when applying defense techniques. For the former, we used a model trained with $\epsilon = 4$. For the latter, we used a model trained with $\epsilon = 4$ and $m = 4$ (that explains why we only run attacks with $\epsilon = 4$). Both models are available in the published repository.

Recommenders. We realized the FM model using the LightFM library [30]. We trained the model for 100 epochs and left all the parameters with the library default values. All the other models were implemented in TensorFlow. As for VBPR and AMR, we trained the models following the training settings adopted by Tang et al. [47], while for DVBPR, we adopted the same parameters found in the official implementation (https://github.com/kang205/DVBPR). On the contrary, we chose ACF hyper-parameters through *grid search* (batch size: [32, 64, 128], learning rate: [0.01, 0.1], regularizer: [0, 0.01, 0.001]). Learning rate and regularizer were set to 0.1 and 0 respectively, while the batch size was set to 32 for `Tradesy` and 64 for `Amazon Women` and `Amazon Men`. The rationale behind the fact that we applied a grid-search to test ACF is that the other VRSs were originally presented and trained in a highly-comparable scenario to ours, i.e., the same datasets, while ACF has been tested by Chen et al. [10] on diverse datasets. At the end of the grid-search, we found that the ACF loss function reaches the convergence after 20 epochs on our tested datasets. For each dataset, we used the *leave-one-out* training-test protocol putting in the test set the last time-aware user's interaction.

4 EXPERIMENTAL RESULTS

In this section, we present and discuss the `VAR` experimental results. As for the recommendation results, we evaluate the top-20 recommendation lists (we indicate *CHR*@20 as *CHR*). In the remainder of this section, we adopt the notation <dataset, VRS, attack, defense> to indicate a specific VAR configuration, where each field

Algorithm 1 Experimental Scenario of `VAR`.

Train the VRS on clean item images.
Measure the *Base* $CHR@K$ for each category C.
Select origin (O) and target (T) categories s.t. $CHR_O@K < CHR_T@K$.
Perform an Adv. Attack against IFE to misclassify O-Images as T.
Poison the dataset with the adversarial perturbed item images.
Measure the $CHR_O@K$ of the O-Products after the Adv. Attack.

Table 2: Averaged origin-target *CHR* evaluated on the `VAR` experimental evaluation in defense-free settings.

Dataset	*Origin*	*CHR*	*Target*	*CHR*	CHR_T/CHR_O
`Amazon Men`	Sandal	0.4508	Running Shoe	2.0191	4.4787
`Amazon Women`	Jersey, T-shirt	0.6324	Brassiere, Bandeau	1.8531	2.9305
`Tradesy`	Suit	0.3810	Trench Coat	1.5371	4.0345

varies depending on the dimensions described in Section 3. The results reported in this section have been computed following the experimental scenario presented in Algorithm 1. Table 2 shows the statistics of the selected origin/target categories.

4.1 Analysis of Attacks and Defenses' Efficacy

This paragraph analyzes the success rate (*SR*) and the feature loss (*FL*) of the adversarial attacks against the IFE components reported in Table 3. Since we did not apply any defensive strategy to the custom DNN adopted for DVBPR (see Section 2.2), the corresponding table cells have been left blank.

4.1.1 Attack Success Rate. Results showed in Table 3 confirm PGD and C&W as the strongest attacks when applied to reduce the classification accuracy of a defense-free CNN classifier. For instance, PGD reaches a near-100% *SR* on `Amazon Men` and 100% *SR* on `Tradesy`, C&W's *SR* is always more than 89%, while FGSM never gets the same results, showing the lowest performance, i.e., 18%, on `Amazon Women`. As expected, this behavior varies with defense strategies. Under this setting, C&W emerges as the best offensive solution against defense strategies, as already demonstrated in [8]. For example, we observe an average *SR* reduction in the *SR* results of 77% for FGSM, 82% for PGD, and 62% for *C&W*.

Hence, we compare the *SR* results to the variation of visual-aware recommendations for the items belonging to the perturbed category of images. Our assumption here is to empirically find a *conformity* between *classification* and *recommendation* metrics on the definition of *successful* attack. Surprisingly, Table 4 shows a different trend from the one observed earlier for the defense-free setting. As far as the *CHR* is concerned, FGSM and C&W attacks are almost aligned on average, i.e., 0.6222 and 0.6212 respectively, but PGD is the best performing attack, i.e., 0.7932 averagely. We also see *discrepancies* under defense-activated scenarios, in which all calculated *CHR* values show negligible differences, with FGSM and C&W mildly outperforming PGD, i.e., especially on AT.

Observation 1. *Attack success rate is not directly related to the effects on the recommendation performance. In other words, being powerful enough to lead a classifier in mislabelling an origin product image towards a target class does not justify the recommendation lists' effects.*

4.1.2 Features Loss. Motivated by the previous observations, we investigate the Feature Loss (*FL*) between original and attacked

Table 3: Average values of Success Rate (*SR*) and Feature Loss (*FL*) for each combination. *FL* values are multiplied by 10^3.

Data	VRS	Att.	Traditional *SR*	Traditional *FL*	Adv. Train. *SR*	Adv. Train. *FL*	Free Adv. Train. *SR*	Free Adv. Train. *FL*
Amazon Men	FM, VBPR, AMR	FGSM	65%	14.0948	18%	0.0330	15%	0.0278
		PGD	**97%**	**36.8843**	18%	0.0334	15%	0.0283
		C&W	89%	20.5172	**48%**	**2.8022**	**42%**	**1.9080**
	ACF	FGSM	65%	9.0480	18%	0.0944	15%	0.0951
		PGD	**97%**	9.2606	18%	0.0944	15%	0.0954
		C&W	89%	**10.4917**	**48%**	**0.7582**	**42%**	**0.4955**
	DVBPR	FGSM	65%	**16.4055**	—	—	—	—
		PGD	**97%**	16.1151	—	—	—	—
		C&W	89%	16.3442	—	—	—	—
Amazon Women	FM, VBPR, AMR	FGSM	18%	9.6677	0%	0.0113	0%	0.0094
		PGD	85%	**27.6645**	0%	0.0119	0%	0.0102
		C&W	**89%**	21.2380	**6%**	**0.1770**	**6%**	**0.3376**
	ACF	FGSM	18%	9.3257	0%	0.0346	0%	0.0424
		PGD	85%	8.3596	0%	0.0352	0%	0.0436
		C&W	**89%**	**11.2079**	**6%**	**0.0399**	**6%**	**0.0594**
	DVBPR	FGSM	18%	20.6968	—	—	—	—
		PGD	85%	17.2065	—	—	—	—
		C&W	**89%**	**24.4750**	—	—	—	—
Tradesy	FM, VBPR, AMR	FGSM	83%	21.4011	43%	0.0308	30%	0.0274
		PGD	**100%**	**53.4589**	43%	0.0311	30%	0.0273
		C&W	**100%**	25.9374	**80%**	**2.1185**	**63%**	**1.9739**
	ACF	FGSM	83%	14.6235	43%	0.0912	30%	0.1069
		PGD	**100%**	10.7754	43%	0.0899	30%	0.1044
		C&W	**100%**	**15.6256**	**80%**	**1.8834**	**63%**	**1.5343**
	DVBPR	FGSM	83%	24.7173	—	—	—	—
		PGD	**100%**	27.0801	—	—	—	—
		C&W	**100%**	**33.6879**	—	—	—	—

samples (as shown in Table 3). The "VRS" column combines the models according to both the IFE and the extraction layer used in the recommendation task. Our assumption here is to empirically find that high distances in the *feature* space correspond to high values of *CHR* and *CnDCG* (we leave the *SR* out of the discussion due to the previous finding). Comparing the results in Tables 3 and 4, we confirm a correlation between the variation of *FL* and the attack efficacy on VRSs. For instance, we see how PGD and C&W higher adversarial power in poisoning the VRS on `Amazon Women`—both on traditional and defensive scenarios— is also evident in the calculated *FL* on the same dataset. Additionally, we notice that the *FL* obtained for DVBPR on `Amazon Women` and `Tradesy` is averagely higher than the one on `Amazon Men`, i.e., 20.7928 and 28.4951 on `Amazon Women` and `Tradesy` respectively *vs.* 16.2883 on `Amazon Men`. We also identify the same trend on DVBPR from a *recommendation* point of view, i.e., there could be an attack method able to increase the *base*-case *CHR*.

Observation 2. *The modification of VRS is closely linked to the magnitude difference between original and perturbed image features. In short, perturbations leading to larger feature modifications may cause a strong influence on the recommendability of the altered item categories.*

4.1.3 Category-based Performance. After having justified the results in Table 4, we discuss the category-based measures across models and datasets studying the *CHR* and *CnDCG*.

The results on FM show that adversarial attacks are always effective in the case of defense-free settings with an across-dataset average *CHR* and *CnDCG* improvements of +5.46% and 6.51%, respectively. Furthermore, the application of the two defenses shows a partial defense. For instance, the <`Amazon Men`, FM, (AT, FAT)> combinations verify that the recommendability of the perturbed category could even receive small negative variations, e.g., an average reduction of *CHR* of -5.94% in the AT case. However, it can be seen that attacks are still effective in any <(`Amazon Women`, `Tradesy`),

Table 4: Results of the `VAR` framework. A *CHR*, or *CnDCG*, higher than the *Base* means that the attack is effective. For each <dataset, VRS, defense> combination we put in bold the most efficient attack.

Data	VRS	Att.	Traditional *CHR*	Traditional *CnDCG*	Adv. Train. *CHR*	Adv. Train. *CnDCG*	Free Adv. Train. *CHR*	Free Adv. Train. *CnDCG*
Amazon Men	FM	*Base*	*0.4960*	*0.0246*	*0.4082*	*0.0204*	*0.4048*	*0.0202*
		FGSM	**0.5309***	**0.0266***	0.3886	0.0198*	0.3821*	0.0194*
		PGD	0.5293*	**0.0266***	0.3795*	0.0193*	0.3811*	0.0193*
		C&W	0.5258*	0.0263*	0.3837*	0.0194*	0.3871*	0.0194*
	VBPR	*Base*	*0.6531*	*0.0293*	*0.3074*	*0.0141*	*0.3775*	*0.0159*
		FGSM	0.5824*	0.0299	0.6164*	0.0323*	0.5860*	0.0283*
		PGD	**1.1480**	**0.0538***	0.6410*	0.0324*	0.5918*	0.0286*
		C&W	0.6132*	0.0290	**0.6880***	**0.0336***	**0.6642***	**0.0348***
	AMR	*Base*	*0.3944*	*0.0196*	*0.5037*	*0.0232*	*0.1076*	*0.0038*
		FGSM	0.3347*	0.0150*	0.4426*	0.0235	0.4178*	0.0187*
		PGD	**0.8365**	**0.0418***	0.4519*	**0.0242**	0.4263*	0.0193*
		C&W	0.3678	0.0170*	0.4371*	0.0230	**0.4451***	**0.0202***
	ACF	*Base*	*0.5574*	*0.0278*	*0.3560*	*0.0176*	*0.3565*	*0.0176*
		FGSM	**0.5692***	**0.0282***	**0.3773***	**0.0185***	0.3517	0.0172*
		PGD	0.5610	0.0280	0.3731*	0.0183*	0.3521	0.0172*
		C&W	0.5628	0.0279	0.3690*	0.0181*	0.3471*	0.0169*
	DVBPR	*Base*	*0.6945*	*0.0359*	—	—	—	—
		FGSM	0.6579*	0.0329*	—	—	—	—
		PGD	0.5549*	0.0281*	—	—	—	—
		C&W	0.6414*	0.0306*	—	—	—	—
Amazon Women	FM	*Base*	*0.6956*	*0.0347*	*0.4720*	*0.0236*	*0.3231*	*0.0162*
		FGSM	0.7030	0.0354*	0.4804*	0.0243*	0.3022*	0.0150*
		PGD	**0.7144**	**0.0356***	**0.4854***	**0.0244***	0.3093*	0.0155*
		C&W	0.6935	0.0346	0.4761*	0.0240	0.2877*	0.0144*
	VBPR	*Base*	*0.4475*	*0.0210*	*0.5213*	*0.0251*	*0.3476*	*0.0161*
		FGSM	0.3933*	0.0182*	0.6199*	0.0310*	0.6204*	0.0318*
		PGD	**0.9530***	**0.0459***	**0.6463***	**0.0327***	**0.6413***	**0.0330***
		C&W	0.4215*	0.0179*	0.6457*	0.0326*	0.5880*	0.0302*
	AMR	*Base*	*0.9907*	*0.0462*	*0.8640*	*0.0454*	*0.5207*	*0.0303*
		FGSM	**1.4178***	**0.0862***	0.7379*	0.0334*	0.4658*	0.0230*
		PGD	1.2720*	0.0713*	0.6664*	0.0307*	0.5003*	0.0250*
		C&W	1.3762*	0.0761*	0.7390*	0.0336*	0.5112*	0.0252*
	ACF	*Base*	*0.9903*	*0.0511*	*0.6890*	*0.0349*	*0.4338*	*0.0219*
		FGSM	0.9895	0.0509	0.6935	0.0350	0.4737*	0.0242*
		PGD	0.9932	0.0512	0.6915	0.0348	0.4759*	**0.0243***
		C&W	**0.9947**	**0.0514***	**0.6943**	**0.0351**	**0.4774***	**0.0243***
	DVBPR	*Base*	*0.7787*	*0.0370*	—	—	—	—
		FGSM	0.7959*	0.0388*	—	—	—	—
		PGD	0.7407	0.0385*	—	—	—	—
		C&W	**0.9002***	**0.0436***	—	—	—	—
Tradesy	FM	*Base*	*0.3424*	*0.0167*	*0.3629*	*0.0183*	*0.4774*	*0.0241*
		FGSM	0.3696*	0.0183*	0.3800*	0.0189	0.5234*	0.0268*
		PGD	0.3664*	0.0180*	0.3661*	0.0181	0.5172*	0.0265*
		C&W	**0.3800***	**0.0190***	**0.3968***	**0.0196***	**0.5236***	**0.0269***
	VBPR	*Base*	*0.4201*	*0.0213*	*0.3011*	*0.0139*	*0.3243*	*0.0146*
		FGSM	0.5313*	0.0293*	**0.5182***	**0.0277***	**0.5770***	**0.0294***
		PGD	**1.3126***	**0.0748***	0.4508*	0.0226*	0.5330*	0.0268*
		C&W	0.4603*	0.0251*	0.4884*	0.0252*	0.5612*	0.0274*
	AMR	*Base*	*0.3710*	*0.0174*	*0.1638*	*0.0065*	*0.2215*	*0.0094*
		FGSM	0.4855	0.0246*	**0.3662***	0.0190*	**0.4094**	**0.0200***
		PGD	**1.0768***	**0.0585***	0.3490*	0.0180*	0.3683*	0.0181*
		C&W	0.4372*	0.0214*	0.3648*	**0.0196***	0.3672*	0.0172*
	ACF	*Base*	*0.3712*	*0.0192*	*0.3685*	*0.0178*	*0.4476*	*0.0218*
		FGSM	**0.3774***	**0.0195***	0.3864*	0.0189*	**0.4606***	**0.0223**
		PGD	0.3728	0.0193	0.3869*	**0.0190***	0.4604*	**0.0223**
		C&W	0.3734	0.0193	**0.3875***	**0.0190***	0.4561*	0.0221
	DVBPR	*Base*	*0.5810*	*0.0298*	—	—	—	—
		FGSM	**0.5956***	**0.0365***	—	—	—	—
		PGD	0.4668*	0.0238*	—	—	—	—
		C&W	0.5701*	0.0308*	—	—	—	—

* *denotes statistically significant results (p-value* ≤ 0.05*).*

AT, FM> scenarios, e.g., $CHR_{PGD} = 0.4854 > CHR_{Base} = 0.4720$ in the `Amazon Women` dataset.

As regards VBPR, PGD is the most impactful strategy in any defense-free setting. For instance, PGD leads to a three times *CHR* increase of the attacked category, i.e., suit, on the `Tradesy` dataset. It means that the adversary has been able to push the class of products in the recommendation lists very effectively, ensuring that a *suit* will be recommended at least one time for each top-20 recommendation list, i.e., $CHR = 1.3126 > 1$ in the <`Tradesy`, VBPR, PGD, T> setting. Additionally, we observe that there are effective attacks in any defended setting.

Observation 3. *The adversarial robustification strategies have not protected VBPR from the injection of perturbed images, although they got high performance in protecting the classification.*

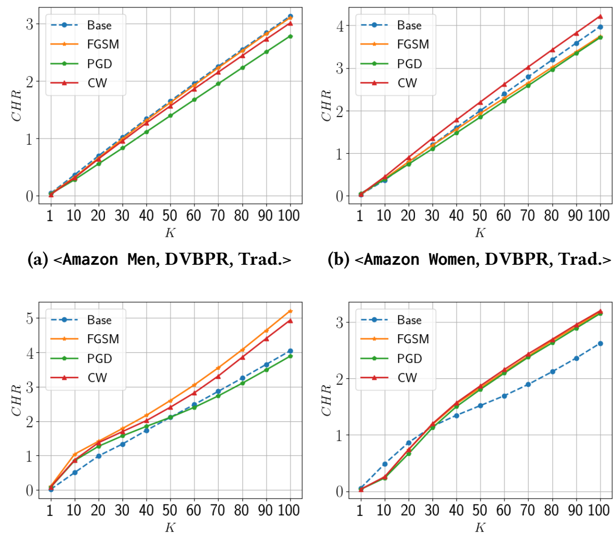
The third tested VRS is AMR. We chose this model since it is the first VRS to **integrate adversarial protection by design**, so we expected to get a limited variation in traditional performance under attack settings. Surprisingly, results show that AMR is prone to the effects of attacks as much as VBPR. For example, the PGD method represents the biggest security threat on the VRS in defense-free settings with an average *CHR* improvements of +48.84% across the three datasets. Moreover, we observe that <AMR, (AT, FAT)> models do not protect the proposed adversarial threat model, notwithstanding the two defense techniques applied on both the IFE and the VRS, respectively. For instance, $CHR = 0.4451 > 0.1076$ when comparing *C&W* and *Base* in <`Amazon Men`, AMR, FAT> experiments. We justify AMR's *low-quality* protection against the tested attacks by the fact that it applies the adversarial regularization directly on the extracted visual features [47], whereas in our experimental framework, the perturbation is produced at the pixel level.

Observation 4. *Combining state-of-the-art adversarial robustification of the IFE, e.g., AT and FAT, and the adversarial robustification of the VRS, e.g., the adversarial regularization of a RS [24]) does not guarantee the protection of the performance.*

The fourth model is ACF. This model is the most robust in the case of defense-free settings when compared with the other models that use the visual features extracted from an external pre-trained IFE, i.e., FM, VBPR, and AMR. Indeed, both *CHR* and *CnDCG* show average variations of +0.79% and 0.61%, respectively, that are much smaller than the one observed in the other models, e.g., the variation is 44.71% in VBPR experiments. The same limited adversary efficacy in altering the recommendation lists can also be seen in the defended settings.

Observation 5. *The tendency of ACF to be naturally robust to the tested attacks can be associated with the fact that it integrates a more semantic-oriented latent representation of the images, e.g., the feature map, and the recommendation task depends not only on the features extracted from the attacked item but also from the set of the items previously voted by each user.*

Finally, we study whether the attacks against a pre-trained CNN used for image classification are transferable to DVBPR, a VRS that learns the deep visual features within the downstream recommendation task. It can be seen that the adversary's efficacy depends on the attacked dataset. Indeed, results in Table 4 shows that DVBPR is not affected by an increase of *CHR* in the `Amazon Men` dataset. However, we can see that C&W effectively varies *CHR* by more than the +10% in the `Amazon Women` dataset and FGSM changes the *CHR* by +2.52% in `Tradesy`.



(a) <`Amazon Men`, DVBPR, Trad.> (b) <`Amazon Women`, DVBPR, Trad.>

(c) <`Amazon Women`, AMR, Trad.> (d) <`Amazon Women`, AMR, Adv. Train.>

Figure 2: Plots of *CHR@K* by varying K from 1 to 100.

Observation 6. *The learning of personalized deep visual representation of product images by DVBPR could be fooled by adversarial attacks transferred from another-trained DNN, raising the need for further investigation to robustify these models.*

4.1.4 Results at Varying Top-k. Before we move to the study of overall recommendation performance, we investigate the effects of adversarial attacks and defenses by varying the length of recommendation lists *K*. Figure 2 reports plots related to two possible interesting cases shown in Table 4: (1) the case where DVBPR was robust, or not, against the tested attacks, and (2) the case where by changing the IFE from a traditional to an adversarial trained one then AMR showed more robust *CHR*@20 results in the `Amazon Women` dataset. The first scenario in Figure 2a shows that the robust behavior of DVBPR observed in the `Amazon Men` dataset is also confirmed on top-100 recommendation lists, while Figure 2b verifies that C&W sill is a powerful strategy to push the perturbed category of product with the difference with the *CHR@K*-baseline that increases with K. Regarding the second set of plots, Figure 2c confirms that FGSM and C&W make the adversarial regularization of the VRS ineffective since the *CHR@K* is always larger than *Base* as k increases, while Figure 2d returns a new unknown phenomenon related to the fact that the robustification of <AMR, AT>, observed on short recommendation lists, e.g., K=20 in Table 4, could be not confirmed on longer recommendation lists, e.g., K=100 $CHR@100_{C\&W} \simeq 1.22 \times CHR@100_{Base}$.

Observation 7. *Adversarial attacks' efficacy might be even more evident when analyzing longer top-K lists raising the need for more powerful defensive strategies in cases where the model is robust on short-length recommendation lists.*

4.2 Overall Recommendation Variations

Table 5 reports the variations of *Rec* and *Nov* measured on an attacked recommenders. The aim is to understand whether the application of defenses adopted to alleviate attacks' influence could generate a drastic variation of the overall recommendation performance. For instance, Δ_{EFD} on AMR has positive values independently from the application of defense mechanisms in the case of `Amazon Men`, i.e., Δ_{EFD} = +14.74% in the case of FAT defense. In contrast, VBPR gets more negative variations across both metrics in the cases tested on the `Amazon Men` dataset. This behavioral pattern is different in the case of `Amazon Women`. Indeed, VBPR measures get positive variation for FAT experimental cases, e.g., Δ_{Rec} = +5.53% on the Traditional model, while negative for the AT one, e.g., $\Delta_{Rec} = -10.51\%$.

Observation 8. *The application of powerful attacks has not tragically worsened the accuracy and beyond accuracy performance. On the contrary, some measures have significantly improved as a consequence of the attack.*

Analyzing the overall variations across the VRS, we observe that ACF and DVBPR are the models less likely to get substantial-overall performance variations when under attacks. For instance, ACF shows a total average variation of -1.22%, while DVBPR by -2.17%. On the contrary, FM, VBPR, and AMR are the models with less stable overall recommendations. For example, VBPR gets overall variations on both metrics higher than −11%, while AMR shows variations close to +9%.

Observation 9. *Both the ACF attentive mechanisms and the DVBPR personalized image features extracted make the recommendation task less subjected to performance variations when the images of a single category of products are perturbed.*

5 RELATED WORK

5.1 Adversarial Machine Learning

ML models have demonstrated vulnerabilities to adversarial attacks [3, 46], i.e., specifically created data samples able to mislead the model despite being highly similar to their clean version. Particularly, great research effort has been put into finding the minimum *visual perturbation* to attack images to fool CNN classifiers. Szegedy et al. [46] formalized the adversarial generation problem by solving a box-constrained L-BFGS. Goodfellow et al. [18] proposed Fast Gradient Sign Method (FGSM), a simple one-shot attack method that uses the sign of the gradient of the loss function. Basic Iterative Method (BIM) [17] and Projected Gradient Descent (PGD) [33] re-adapted FGSM to create stronger attacks by *iteratively* updating the adversarial perturbation. Carlini and Wagner [8] improved the problem definition presented in [46] and built powerful attacks in deceiving several detection strategies [7]. Along with the proposed attacks, many solutions have also been provided regarding defense. Adversarial Training [18] creates new adversarial samples at training time, making the model more robust to such perturbed inputs. Defensive Distillation [39] transfers knowledge between two networks to reduce the resilience to adversarial samples but was proven not to be as secure as expected against C & W attacks [6]. Free Adversarial Training [45] truly eases the computational complexity of adversarial training.

Table 5: Overall recommendation variations results (Δ_{Rec} and Δ_{Nov} reported for `VAR`).

Data	VRS	Traditional Δ_{Rec}	Traditional Δ_{EFD}	Adv. Train. Δ_{Rec}	Adv. Train. Δ_{EFD}	Free Adv. Train. Δ_{Rec}	Free Adv. Train. Δ_{EFD}
`Amazon Men`	FM	+8.00	+38.45	-30.08	-18.04	-4.52	-4.17
	VBPR	+2.37	-1.33	-45.49	-41.58	-31.42	-33.76
	AMR	+0.75	+1.37	+5.92	+14.74	+2.50	+9.97
	ACF	-1.54	-4.02	-0.69	+0.35	+6.19	0.00
	DVBPR	+6.17	+4.72	—	—	—	—
`Amazon Women`	FM	+8.42	+0.81	+23.69	+20.82	+9.02	+9.59
	VBPR	-1.74	-0.95	-10.51	-13.47	+1.29	+3.39
	AMR	-0.26	-1.39	+6.04	+5.71	+5.34	+3.90
	ACF	-1.96	-1.74	+1.72	-4.32	+5.50	+10.95
	DVBPR	-0.24	+2.94	—	—	—	—
`Tradesy`	FM	+5.23	-0.23	+8.51	+11.01	+36.59	+27.7
	VBPR	+2.95	-0.51	+4.50	-4.71	-1.17	-9.85
	AMR	+17.92	+20.88	+24.82	+28.98	+3.48	-2.38
	ACF	-2.38	-2.20	-6.17	-15.55	-4.95	-11.00
	DVBPR	-11.11	-15.47	—	—	—	—

5.2 Visual-based Recommender Systems

The integration of image features in user's preference predictor leads to enhancing both recommendation [11, 22, 23, 36, 56] and search [27, 50, 56] tasks. The intuition is that the visual appearance of product images influences customer's decisions, e.g., a customer who loves red will likely buy red clothes [19]. For instance, He and McAuley [23] extended BPR-MF [42] by integrating high-level features extracted from a pre-trained CNN, while Kang et al. [26] trained the same model in an end-to-end manner by stacking a custom CNN at the top, whose purpose is feature representation learning and not simply classification. Yu et al. [52] added aesthetic information in the recommendation framework to enhance CNNs' extracted features, which carry only semantic content. Yin et al. [51] proposed to incorporate visual features to learn item-to-item compatibility relations for outfit recommendation. Furthermore, Niu et al. [36] injected the visual features into a neural personalized model, and Chen et al. [10] integrated component-level image features, e.g., regions in an image, to learn users' preferences from more informative image representations. In this work, we focused on VRSs that integrate both features extracted from both CNNs pre-trained for a *classification task*, e.g., [10, 22, 36, 47], and CNNs learned within the VRS [26].

5.3 Security of Recommender Systems

Recommender models have been demonstrated to be steadily under security risks. The security of RSs relates to the study of different hand-engineered strategies to generate shilling profiles, which lead to the alteration of collaborative recommendations [31], and their defense mechanisms, e.g., detection [2] and robustness [37]. On the other hand, the application of AML in RSs differs from previous works in the use of optimized perturbations and their respective defenses, which lead to drastic performance reduction [1, 4, 24, 32, 47, 48]. For example, He et al. [24], Yuan et al. [53] used an adversarial training procedure to make the model robust to such perturbations. Furthermore, Tang et al. [47] applied this defense to make the proposed VRS, i.e., AMR, more robust to adversarial perturbations on image features. However, Di Noia et al. [15] noticed the partial protection of VBPR and AMR against targeted adversarial attacks on product images. Recently Cohen et al. [12] proposed a black-box attack strategy to push a target item to higher recommendation positions. Differently from our work, the authors perturbed the product images at inference time, we investigated in VAR the training time insertion of adversarially perturbed product images.

6 CONCLUSION AND FUTURE WORK

We have presented an evaluation framework, i.e., Visual Adversarial Recommendation (VAR), to investigate the effectiveness of robustification mechanisms on the DNNs, i.e., Adversarial Training/Free Adversarial Training, used in VRSs. We have tested three state-of-the-art white-box attacks, i.e., FGSM, PGD, and C&W, to perturb the products' low-recommended product images category. The goal of the studied adversary threat model is to make these pictures misclassified by the DNN toward the class of top-rated products to push their recommendability. Experimental results have shown that defense mechanisms do not guarantee the protections of VRSs against attacks. Interestingly, we have found that *the effectiveness of attacks in altering the recommenders is more related to high feature losses than high success rates*. Additionally, we have also observed that DVBPR, a VRS that learns deep image representations without using external DNNs, is not robust to adversarial samples transferred by attacking other networks. Finally, we have verified that overall recommendation performance has not worsened under the experimented threat model, and defended IFEs may even improve in non-attack settings. These findings raise the need to develop novel defense approaches to protect visually-aware recommender models. The investigation of the reasons behind the models' weakness could get benefit in defending a VRS, and verify whether other multimedia recommenders, e.g., music recommenders, could be affected by the same treats, e.g., push an artist.

REFERENCES

[1] Vito Walter Anelli, Alejandro Bellogin, Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. MSAP: Multi-Step Adversarial Perturbations on Recommender Systems Embeddings. In *The 34th International FLAIRS Conference*. The Florida AI Research Society (FLAIRS), AAAI Press, 1–6. http://sisinflab.poliba.it/publications/2021/ABDDM21

[2] Runa Bhaumik, Chad Williams, Bamshad Mobasher, and Robin Burke. 2006. Securing collaborative filtering against malicious attacks through anomaly detection. In *ITWP 2006*.

[3] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion Attacks against Machine Learning at Test Time. In *ECML-PKDD 2013*.

[4] Yuanjiang Cao, Xiaocong Chen, Lina Yao, Xianzhi Wang, and Wei Emma Zhang. 2020. Adversarial Attacks and Detection on Reinforcement Learning-Based Interactive Recommender Systems. In *SIGIR*. ACM, 1669–1672.

[5] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. 2019. On Evaluating Adversarial Robustness. *CoRR 2019* (2019).

[6] Nicholas Carlini and David A. Wagner. 2016. Defensive Distillation is Not Robust to Adversarial Examples. *CoRR 2016* (2016).

[7] Nicholas Carlini and David A. Wagner. 2017. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In *AISec@CCS 2017*.

[8] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *SP 2017*.

[9] Huiyuan Chen and Jing Li. 2019. Adversarial tensor factorization for context-aware recommendation. In *RecSys 2019*.

[10] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *SIGIR*. ACM.

[11] Xiaoya Chong, Qing Li, Howard Leung, Qianhui Men, and Xianjin Chao. 2020. Hierarchical Visual-aware Minimax Ranking Based on Co-purchase Data for Personalized Recommendation. In *WWW 2020*.

[12] Rami Cohen, Oren Sar Shalom, Dietmar Jannach, and Amihood Amir. 2020. A Black-Box Attack Model for Visually-Aware Recommender Systems. arXiv:2011.02701 [cs.LG] to appear in WSDM 2021.

[13] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–38.

[14] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender Systems Leveraging Multimedia Content. *ACM Comput. Surv.* 53, 5 (2020), 106:1–106:38. https://doi.org/10.1145/3407190

[15] Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2020. TAaMR: Targeted Adversarial Attack against Multimedia Recommender Systems. In *DSN–DSML 2020*.

[16] Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, and Evangelos E. Papalexakis. 2020. All You Need Is Low (Rank): Defending Against Adversarial Attacks on Graphs. In *WSDM 2020*.

[17] Alexey Kurakand Ian J. Goodfellow and Samy Bengio. 2017. Adversarial examples the physical world. In *ICLR 2017*.

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR 2015*.

[19] Kristen Grauman. 2020. Computer Vision for Fashion: From Individual Recommendations to World-wide Trends. In *WSDM 2020*.

[20] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. 2018. Countering Adversarial Images using Input Transformations. In *ICLR 2018*.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016*.

[22] Ruining He and Julian J. McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *WWW 2016*.

[23] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *AAAI 2016*.

[24] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial Personalized Ranking for Recommendation. In *SIGIR 2018*.

[25] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* (2012).

[26] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian J. McAuley. [n.d.]. Visually-Aware Fashion Recommendation and Design with Generative Image Models. In *ICDM 2017*.

[27] Saeid Balaneshin Kordan and Alexander Kotov. 2018. Deep Neural Architecture for Multi-Modal Retrieval based on Joint Embedding Space for Text and Images. In *WSDM 2018*.

[28] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8 (2009), 30–37.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NeurIPS 2012*.

[30] Maciej Kula. 2015. Metadata Embeddings for User and Item Cold-start Recommendations. In *CBRecSys@RecSys 2015*.

[31] Shyong K. Lam and John Riedl. 2004. Shilling recommender systems for fun and profit. In *WWW 2004*.

[32] Yang Liu, Xianzhuo Xia, Liang Chen, Xiangnan He, Carl Yang, and Zibin Zheng. 2020. Certifiable Robustness to Discrete Adversarial Perturbations for Factorization Machines. In *SIGIR*. ACM, 419–428.

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR 2018*.

[34] Jarana Manotumruksa and Emine Yilmaz. 2020. Sequential-based Adversarial Optimisation for Personalised Top-N Item Recommendation. In *SIGIR*. ACM, 2045–2048.

[35] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *SIGIR 2015*.

[36] Wei Niu, James Caverlee, and Haokai Lu. 2018. Neural Personalized Ranking for Image Recommendation. In *WSDM 2018*.

[37] Michael P. O'Mahony, Neil J. Hurley, Nicholas Kushmerick, and Guenole C. M. Silvestre. 2004. Collaborative recommendation: A robustness analysis. *ACM Trans. Internet Techn.* (2004).

[38] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakand Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-LJuang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. 2018. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. *Corr 2018* (2018).

[39] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *SP 2016*.

[40] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS 2015*.

[41] Steffen Rendle. 2010. Factorization Machines. In *ICDM 2010*.

[42] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 209. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009*.

[43] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *WWW*. ACM.

[44] Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 2015. *Recommender Systems Handbook*. Springer.

[45] Ali Shafahi, Mahyar Najibi, AmGhiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, GavTaylor, and Tom Goldstein. 2019. Adversarial training for free!. In *NeurIPS 2019*.

[46] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR 2014*.

[47] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2020. Adversarial Training Towards Robust Multimedia Recommender System. *IEEE Trans. Knowl. Data Eng.* 32, 5 (2020), 855–867.

[48] Xianfeng Tang, Yandong Li, Yiwei Sun, Huaxiu Yao, Prasenjit Mitra, and Suhang Wang. 2020. Transferring Robustness for Graph Neural Network Against Poisoning Attacks. In *WSDM 2020*.

[49] Saúl Vargas. 2014. Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *SIGIR*. ACM, 1281.

[50] Zhijing Wu, Yiqun Liu, Qianfan Zhang, Kailu Wu, Min Zhang, and Shaoping Ma. 2019. The Influence of Image Search Intents on User Behavior and Satisfaction. In *WSDM 2019*.

[51] Ruiping Yin, Kan Li, Jie Lu, and Guangquan Zhang. 2019. Enhancing Fashion Recommendation with Visual Compatibility Relationship. In *WWW 2019*. 3434–3440.

[52] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based Clothing Recommendation. In *WWW 2018*, Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.).

[53] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. Adversarial Collaborative Neural Network for Robust Recommendation. In *SIGIR*. ACM, 1065–1068.

[54] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Trans. Neural Networks Learn. Syst.* 30, 9 (2019), 2805–2824.

[55] Zhenlong Yuan, Yongqiang Lu, Zhaoguo Wang, and Yibo Xue. 2014. Droid-Sec: deep learning android malware detection. In *SIGCOMM 2014*.

[56] YZhang and James Caverlee. 2019. Instagrammers, Fashionistas, and Me: Recurrent Fashion Recommendation with Implicit Visual Influence. In *CIKM 2019*.